Method and Device for Analyzing a Spoken Sequence of Numbers

BACKGROUND OF THE INVENTION

Technical Field

The invention relates to a method and a device for analyzing a spoken sequence of numbers.

Discussion of the Prior Art

A lot of technical applications require recognition of a spoken sequence of numbers. Many mobile telephones comprise the feature of voice dialing by uttering a telephone number. Moreover, electronic commerce applications require the recognition of spoken order numbers and spoken credit card numbers.

WO-A-89 04035 discloses a method for recognizing a number like a telephone number consisting of a plurality of digits. The digits are uttered singly or in sequences. Two utterances comprising one or more digits may be separated by the user-defined placement of pauses. A pause time between two utterances is monitored and when an utterance is followed by a pre-determined pause time interval, the recognized digits will be replied via a speech synthesizer. A further utterance comprising one or more digits can then be started, and only the next utterance will be replied after a subsequent pause.

While recognition of spoken digits and spoken digit sequences works reliably also under adverse noise conditions, automatic recognition of naturally spoken numbers like "twenty two" or "five hundred thirty" is more difficult. This is due to the fact that spoken sequences of numbers like "twenty two" or "five hundred thirty" can stand for more than one numerical value. The spoken sequence of numbers "twenty two", for example, can stand either for the single numerical value "22" or for the two numerical values "20" and "2". As another example, the sequence "five hundred thirty" can stand both for the numerical value "530" or for the two numerical values "500" and "30".

When automatically recognizing a spoken sequence of numbers, the recognition process becomes increasingly difficult if numbers with a large numerical value or a large sequence of numbers have to be analyzed. Thus, the spoken sequence of numbers "thousand four hundred fifty six" can stand for a single numerical value or for up to five numerical values. Altogether, there exist eight possibilities: "1456", "1000" and "4" and "100" and "50" and "6", "1000" and "456", "1000" and "400" and "56", "1000" and "400" and "50" and "6", "1400" and "56", "1400" and "50" and "6", "1450" and "6".

These ambiguities do not only occur in the English language. In the German language , for example, the naturally spoken sequence of numbers "einhundert zehn" can stand both for the single numerical value "110" and the two numerical values "100" and "10". However, the ambiguities relating to the one or more numerical values of a spoken sequence of numbers may be different in different languages. While e. g. in the French language "quarante sept" can stand for both the single numerical value "47" or the two numerical values "40" and "7", this ambiguity does not occur in the German language. In the German language the numerical value "47" is spoken as "siebenundvierzig" and the sequence of the two numerical values "40" and "7" is spoken as "vierzig sieben".

There is, therefore, a need for a method and device for analyzing a spoken sequence of numbers which allow a robust distinction between different semantic interpretations with respect to the one or more numerical values comprised therein.

SUMMARY OF THE INVENTION

The present invention satisfies this need by providing a method for analyzing a spoken sequence of numbers, wherein the numbers are recognized by automatic speech recognition and wherein the method comprises determining a pause length between two consecutive numbers and deciding whether or not the two consecutive numbers belong to a single numerical value on the basis of the determined pause length. A device for analyzing a spoken sequence of numbers comprises an automatic speech recognizer, a prosodic unit for determining a pause length between two consecutive numbers and a processing unit for deciding whether or not the two consecutive numbers belong to a single numerical value on the basis of the determined pause length.

According to the invention, the speaking pause length between two consecutively spoken numbers is used as the single prosodic criterion or as one of a plurality of prosodic criteria for assessing whether or not the two consecutively spoken numbers belong to a single numerical value or to two different numerical values. The speaking pause length is a robust prosodic criterion for analyzing a spoken sequence of numbers. Further prosodic parameters apart from the speaking pause length on which the decision whether or not two consecutively spoken numbers belong to a single numerical value can be based are known from E. Nöth et al "Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen", in Paulus, Wahl (ed.), Mustererkennung 1997, Informatik aktuell, Springer-Verlag, Heidelberg, 1997, pages 37-52, herewith incorporated by reference.

The decision whether or not two consecutively spoken numbers belong to a single numerical value can be a "hard" decision or a "soft" decision. The "hard" decision can be based on determining whether or not certain thresholds of prosodic parameters have been exceeded. A "soft" decision may be made by means of a so-called classifier, e.g. a neuronal network, which takes into account a plurality of prosodic parameters and which produces e.g. a propability decision.

According to a preferred embodiment of the invention, it is automatically decided that two consecutive numbers do not belong to a single numerical value if a certain pause length threshold is exceeded. Such a mechanism corresponds to the acoustical perception of a human listener. The two spoken numbers "20" and "2" e. g. will clearly be perceived by the human listener as two separate numerical values (i. e. "20" and "2") if a speaking pause of sufficient duration is made between speaking the numbers "20" and "2". On the other hand, the spoken numbers "20" and "2" will be perceived as a single numerical value (i. e. "22") if no or almost no speaking pause is made.

The speaking pause length threshold which forms the basis for the decision whether or not two consecutive numbers belong to a single numerical value can initially be set to a certain value. This value can be an empirical value estimated on the basis of a representative speech database. The pause length threshold can also be adjustable. This allows a user to adapt the speaking pause length threshold to his own manner-of-speaking, e. g. by changing the threshold value in system settings of the device.

It has been found that robust setting of a pause length threshold is strongly interrelated with speech tempo which in turn depends on the individual speaker. In reality, the speech tempo of different speakers can vary within a wide range. According to a preferred embodiment of the invention, the pause length threshold is therefore automatically adapted to the current user's speaking habit. This can e. g. be done by analyzing previously determined speaking pause lengths within one or more previously uttered numerical values which the user has already acknowledged to be correct. A new pause length threshold can then either be set to the mean or the median computed over these previously determined speaking pause lengths or it can be set anywhere between the old threshold and the mean or median value of the previously determined speaking pause lengths. In other words: the pause length threshold is shifted.

The decision whether or not two consecutively spoken numbers belong to a single numerical value can be made more robust if the decision is not only based on the speaking pause length but also on the previously mentioned further prosodic parameters apart from the speaking pause length. These further prosodic parameters can relate to a phoneme duration like phrase-final lengthening or pre-boundary lengthening, the shape of the energy contour or specific pitch movements like phrase-final fall. Preferably, respective thresholds are also provided for these further prosodic parameters. The decision whether or not two consecutive numbers belong to a single numerical value can accordingly also be based on the criterion whether or not a respective threshold of a further prosodic parameter has been exceeded.

Like the pause length threshold, the respective thresholds of further prosodic parameters can be user-adjustable or be automatically adjusted dependent on the user's speaking habit or be adjusted in accordance with appropriate training data. Moreover, previously determined further prosodic parameters of previously uttered numerical values which the user has already acknowledged to be correct can be used for shifting respective thresholds of the prosodic parameters.

In many languages, connecting words between two consecutive numbers of a spoken sequence of numbers indicate that the two consecutive numbers belong to one numerical value. In the English language, e. g., such a connecting word is the word "and". Thus, the spoken sequence of numbers "one hundred and ten" usually stands for the numerical value "110", even if the total pause length between "hundred" and "ten", the pause length between "hundred" and "and" or the pause length between "and" and "ten" exceeds a previously set pause length threshold.

In order to correctly analyze a spoken sequence of numbers comprising one or more connecting words between two consecutive numbers, a preferred embodiment of the invention comprises the feature of recognizing such a connecting word. According to a first variant of the invention, it is determined that two consecutive numbers belong to a single numerical value every time a connecting word is arranged between the two numbers.

According to a second variant, upon recognition of a connecting word between two consecutive numbers, the pause length threshold for determining whether or not the two consecutive numbers belong to a single numerical value is changed. In other words: upon recognition of a connecting word, the decision whether or not two consecutive numbers belong to a single numerical value is based on a different pause length threshold as in case no such connecting word is recognized. Consequently, two different pause length thresholds are utilized. Analyzing a spoken sequence of numbers thus becomes more robust because in certain cases the consecutive numbers belong to different numerical values although a connecting word is arranged therebetween, especially in cases where the pause length between the two consecutive numbers is extremely long (e. g. when a user places long pauses between the connecting word and the number preceding or following the connecting word).

There exist several possibilities for determining a speaking pause length between two consecutive numbers of a spoken sequence of numbers. The pause length can e. g. be directly determined by measuring a silence interval between two consecutively spoken numbers. This can be done with a so-called voice activity detector. A speaking pause length can also be determined indirectly using the information obtained as a byproduct from the process of automatic speech recognition. During automatic speech recognition not only the words themselves but also their respective start and end points on a time axis are computed. The pause length can thus be determined based on an end point of the first of two consecutive numbers and a starting point of a second of two consecutive numbers. Especially in noisy environments, this technique usually leads to more robust results than measuring a silence interval between two consecutive numbers.

BRIEF DESCRIPTION OF THE DRAWINGS

Further aspects and advantages of the invention will become apparent upon reading the following detailed description of preferred embodiments of the invention and upon reference to the drawings in which:

Fig. 1 is a schematic diagram of a device for analyzing a spoken sequence of numbers according to the invention; and

Fig. 2 is a schematic diagram of a method for analyzing a spoken sequence of numbers according to the invention.

DESCRIPTION OF THE PREFERERD EMBODIMENTS

In Fig. 1, a schematic diagram of a device 100 for analyzing a spoken sequence of numbers according to the invention is illustrated. The analyzing device 100 depicted in Fig. 1 comprises an automatic speech recognizer 120, a prosodic unit 140 for determining a pause length between two consecutive numbers, a processing unit 160 for deciding if the two consecutive numbers belong to a single numerical value and an input unit 180.

Upon speaking a sequence of numbers like "five hundred thirty", the automatic speech recognizer 120 recognizes each of the spoken numbers as well as connecting words comprised within the spoken sequence of numbers. During the recognition process, the starting and end points in time of the recognized numbers and connecting words are computed. These starting and end points are output to the prosodic unit 140 for determining the pause length between two consecutive numbers or between a connecting word and a preceding or subsequent number.

The processing unit 160 receives input from both the automatic speech recognizer 120 and the prosodic unit 140. Based on the numbers recognized by the automatic speech recognizer 120, the presence of connecting words between two consecutive numbers and the pause length between two consecutive numbers or a connecting word and a number preceding or following the connecting word, the processing unit 160 analyzes the spoken sequence of numbers with respect to the one or more numerical values contained therein.

The processing unit 160 decides whether or not two consecutive numbers belong to a single numerical value on the basis of a pause length threshold. This pause length threshold is initially set to a value between 100 ms and 1 s, preferably to a value of 200 ms.

By means of an input unit 180 a user has the possibility to adapt this initial threshold to his own manner-of-speaking. The input unit 180 comprises a graphical or physical slide bar allowing to adjust the threshold within a predetermined range. The input unit 180 also allows selection of an automatic adaptation of the threshold to the speaking habit of one or more users of the device 100.

The function of the device 100 is hereinafter described in more detail with reference to Fig. 2.

First of all, a pause length threshold Θ is set automatically or by the user or according to appropriate training data to a certain value. Then, the user speaks the sequence "five hundred thirty" consisting of the three numbers "five", "hundred" and "thirty". These spoken numbers are subjected to automatic speech recognition in the automatic recognizer 120. The automatic speech recognizer 120 recognizes the three numbers "five", "hundred" and "thirty" with their respective starting and end points. The detection of the respective starting and end points indicates that there is a first pause between the first number "five" and the second number "hundred" and a second pause between the second number "hundred" and the third number "thirty".

The starting and end points of the three numbers are input to the prosodic unit 140 which determines a pause length P1 of the first pause as well as a pause length P2 of the second pause. The three numbers recognized by the automatic speech recognizer 120 and the two pause lengths P1 and P2 determined by the prosodic unit 140 are input to the processing unit 160 which decides if two consecutive numbers belong to a single numerical value on the basis of the measured pause lengths P1 and P2.

If both the pause length P1 and the pause length P2 exceed the pause length threshold Θ, the processing unit 160 decides that the spoken sequence of numbers contains three numerical values, i. e. "5", "100" and "30". If neither of the two pause lengths P1 and P2 exceeds the pause length threshold Θ, the processing unit 160 decides that the spoken sequence of numbers contains a single numerical value, i. e. "530".

If the processing unit 160 determines that only the first pause length P1 exceeds the pause length threshold Θ, it decides that the spoken sequence of numbers contains the two numerical values "5" and "130". On the other hand, if only the second pause length P2 exceeds the pause length threshold Θ, the processing unit 160 decides that the spoken sequence of numbers contains the two numerical values "500" and "30".

According to the method depicted in Fig. 2, the pause length P1 is determined prior to the pause length P2. This allows to analyze the spoken sequence of numbers in the order the numbers are spoken. Of course, the pause lengths P1 and P2 may also be determined and analyzed in a different order. This may necessitate that all numbers of the sequence of numbers have to be spoken prior to the analyzing step.

Although the method depicted in Fig. 2 relates to a decision which is solely based on the determined pause length, the prosodic unit 140 depicted in Fig. 1 may also determine further prosodic parameters apart from the pause length and the decision may also be based on these further prosodic parameters. Besides, the automatic speech recognizer 120 may also recognize connecting words within a spoken sequence of numbers and the processing unit 160 may, upon recognition of a connecting word, apply a different threshold regarding the one or more prosodic parameters on which the decision is based. Also, the decision can be based solely on one or more prosodic parameters apart from the pause length.

The device 100 and the method according to the invention may be used for many applications, e. g. stationary electronic commerce systems or mobile applications like mobile telephones.